Predicting Radiologists' Gaze with Computational Saliency Models in Mammogram Reading

Jianxun Lou, Hanhe Lin, Philippa Young, Richard White, Zelei Yang, Susan Shelmerdine, David Marshall, Emiliano Spezi, Marco Palombo and Hantao Liu

***Abstract*—Previous studies have shown that there is a strong correlation between radiologists' diagnoses and their gaze when reading medical images. The extent to which gaze is attracted by content in a visual scene can be characterised as visual saliency. There is a potential for the use of visual saliency in computer-aided diagnosis in radiology. However, little is known about what methods are effective for diagnostic images, and how these methods could be adapted to address specific applications in diagnostic imaging. In this study, we investigate 20 state-of-the-art saliency models including 10 traditional models and 10 deep learning-based models in predicting radiologists' visual attention while reading 196 mammograms. We found that deep learning-based models represent the most effective type of methods for predicting radiologists' gaze in mammogram reading; and that the performance of these saliency models can be significantly improved by transfer learning. In particular, an enhanced model can be achieved by pre-training the model on a large-scale natural image saliency dataset and then fine-tuning it on the target medical image dataset. In addition, based on a systematic selection of backbone networks and network architectures, we proposed a parallel multi-stream encoded model which outperforms the state-of-the-art approaches for predicting saliency of mammograms.**

***Index Terms*—Saliency, Radiology, Mammograms, Deep learning, Transfer learning**

I. Introduction

In the human visual system (HVS), foveal vision covers a small central part of the visual field and provides the most detailed and informative visual signal [1]. The visual attention mechanism drives the foveal vision to prioritise visual stimuli to reduce the consumption of the cerebral cortex. In image perception, visual saliency reflects the extent to which the content in a visual scene attracts gaze, which is an important feature of the HVS. Previous studies have demonstrated that radiologists' diagnoses are strongly related to their gaze when reading medical images [2]–[4]. Being able to predict radiologists' gaze in the form of computational algorithms is beneficial in terms of developing tools for assisting diagnosis [5] and artificial intelligence (AI) models for diagnostic medical imaging [6].

In the literature, visual saliency has been exploited in various medical diagnostic tasks and achieved promising clinical outcomes. Banerjee *et al.* [7] developed a visual saliency-based algorithm for automated brain tumor detection and segmentation in magnetic resonance (MR) images. Bernal *et al.* [8] developed a saliency model for polyp localisation in colonoscopy videos and the model achieved similar performance to clinicians in polyps searching. Yuan *et al.* [9] also used a unified bottom-up and top-down saliency approach to automatically detect polyp regions. In radiology practice, one of the key factors that affects diagnostic accuracy is fatigue [10]. Fatigue can reduce cognitive ability and attention lapses, decrease vigilance, and change gaze patterns of radiologists [10], [11]. In the event of fatigue, the computer generated saliency that represents the standard image reading can be used to assist the radiologist during diagnosis to avoid potential errors caused by fatigue. In addition, since the model simulates the visual attention of radiologists during diagnostic reading, the predicted visual saliency maps can be used as a tool in training of radiology students/trainees to provide them with guidance, feedback, and reflection [12]. It should be noted that the saliency prediction models and many other artificial intelligence (AI) applications in radiology have certain limitations. They are built based on the observers' annotations; and they can follow general structures but lack the creativity e.g., the critical judgement of one's own original ideas. To optimise the integration of AI into medical imaging, research has been undertaken to make AI models more usable and explainable, providing complementary information/intelligence to support radiologists' decision-making [13]–[15]. These studies indicate that automatically generating accurate saliency maps that represent where viewers look in diagnostic images lies at the heart of advanced methods for medical imaging. However, there is a paucity of literature on the analysis of existing saliency prediction methods on diagnostic images, and how these methods could be adapted to best address specific applications in diagnostic imaging.

Existing computational saliency models can be categorized into two classes including traditional and deep learning-based models. Traditional models [16]–[25] apply low-level visual features such as colour, luminance, texture, and contrast, to simulate the visually salient regions in the scene. Rather

Jianxun Lou, David Marshall, Marco Palombo and Hantao Liu are with the School of Computer Science and Informatics, Cardiff University, CF24 4AG Cardiff, United Kingdom.

Hanhe Lin is with the School of Science and Engineering, University of Dundee, DD1 4HN Dundee, United Kingdom.

Philippa Young is with the Breast Test Wales, National Health Service, Cardiff, United Kingdom.

Richard White and Zelei Yang are with the Department of Radiology, University Hospital of Wales, Cardiff, United Kingdom.

Susan Shelmerdine is with the Dept of Clinical Radiology, Great Ormond Street Hospital, WC1N 3JH London, United Kingdom.

Emiliano Spezi is with the School of Engineering, Cardiff University, CF24 3AA Cardiff, United Kingdom.

Marco Palombo is also with Cardiff University Brain Research Imaging Centre, School of Psychology, Cardiff University, CF24 4HQ Cardiff, United Kingdom.

Jianxun Lou is funded by the China Scholarship Council – ID 202008220129.

than designing handcrafted features, deep learning-based models [26]–[34] typically use deep convolutional networks to automatically extract representations from images for saliency prediction. Although there are various saliency prediction algorithms in the literature, most of them focus on predicting the so-called task-agnostic visual attention that mimics gaze during free viewing of natural images. Different from task-agnostic visual attention which is stimulus-driven, the task-specific visual attention is driven by a specific visual task and related to the viewer's prior experience. Visual saliency that is modelled by task-agnostic saliency features may be modulated towards a task-specific model using task-related factors [35]. When radiologists read medical images for a diagnostic purpose, saliency is considered to be determined by the combination of both task-agnostic (i.e., image content) and task-specific (i.e., the diagnostic task and radiologists' prior knowledge) factors [36]. It is not fully known whether visual saliency prediction models that are designed or trained with task-agnostic saliency features of natural scenes still hold their predictive capabilities for medical images, and if so, to what extent. More importantly, it is critical to find ways to develop effective saliency models for diagnostic imaging. Jampani *et al.* [37] studied the relevance of three traditional saliency models in chest X-ray and retinal images in the context of abnormality detection. Wen *et al.* [38] investigated 16 traditional saliency models in three types of medical imaging modalities (including chest computed tomography, chest X-ray images, and whole-body positron emission tomography) relative to natural scenes. These studies suggest that some task-agnostic saliency models developed for natural images are potentially beneficial for medical image saliency prediction. However, there are limitations in these studies. First, the image samples are rather limited, e.g., the dataset in [37] includes 17 chest X-rays and 48 retinal images, and the dataset in [38] consists of only 10 images per modality. Second, the state-of-the-art deep learning-based saliency models were not included in these studies.

Nevertheless, deep learning-based models have achieved promising results in many computer vision applications [39]–[42]. There is a potential to apply transfer learning techniques [43], [44], namely reusing or transfering information from previously learned tasks of natural images for the learning of new tasks of medical images. For example, in image diagnostics, deep learning-based models are generally first trained on large-scale task-unrelated datasets (usually natural image datasets, such as ImageNet [45] and PLACE [46]), then these models are fine-tuned on small-scale medical image datasets related to the target task [47]. In this approach, models are expected to first learn low-level features such as lines, luminance, and contrast in the pre-training phase, and then learn diagnosis-related features in the fine-tuning phase. This approach can be potentially used to learn a saliency prediction model for diagnostic imaging. This is to train models on a large-scale saliency dataset (originated from natural images) before fine-tuning them on medical image datasets. Since medical image datasets are usually limited in size, this approach has the potential to significantly enhance the sample efficiency of a learning model. However, the feasibility and effectiveness of this approach in predicting the saliency of medical images has not been well studied.

The main contributions of this paper are detailed as follows:

1) We conduct an exhaustive comparative study towards a plausible modelling paradigm for mammogram saliency prediction. We study 20 saliency models developed for natural images, including 10 traditional models and 10 deep learning-based models, in the context of screening mammography.
2) We investigate the use of transfer learning for medical image saliency prediction, conducting experiments to fine-tune deep learning-based saliency models on a mammogram saliency dataset, and to pre-train saliency models on a large-scale natural image saliency dataset for predicting mammogram saliency.
3) By harnessing a systematic analysis of different network structures and pre-training strategies, we develop a parallel multi-stream encoded model that gives superior performance on predicting saliency of mammograms.

II. Materials and Methods

A. Eye-tracking mammography dataset

Our study is based on a large-scale mammography eye movement dataset [48]. The original image set consists of 196 mediolateral oblique (MLO) view mammogram images that were extracted from 98 anonymous cases from the University Hospitals KU Leuven in Belgium, each with one MLO view of the left breast and one MLO view of the right breast. An eye-tracking experiment was conducted in a mammography reading room at Breast Test Wales (BTW), a centre participating in the National Health Service Breast Screening Programme (NHSBSP), Cardiff, United Kingdom. The details of the experiment, including the procedure of collecting and processing the eye movement data, can be found in [49]. The eye-tracking data contains the gaze positions of 10 radiologists during the three seconds (corresponding to the viewing time in real practice) of reading these mammograms.

As illustrated in Fig. 1, fixations collected via eye-tracking are graphically represented by a binary fixation map for each mammogram image, where the locations of fixations are rendered as white (grayscale value of 255) pixels and un-fixated areas as black (grayscale value of 0) pixels. To construct a saliency map, each fixation location is expanded by a greyscale patch that is modelled as a Gaussian distribution [48]. The width of the Gaussian distribution represents 2 degrees of visual angle to simulate the size of the fovea [50]. Essentially, the salient regions (i.e., regions with higher density of fixations) represent the locations where human observers focus their gaze with a higher frequency.

B. Computational saliency models

In computer vision, the so-called saliency models are developed with the aim to automatically predict where people look in images. The output of these models is a computed saliency map – a topographic representation indicating conspicuousness of scene locations [51]. As mentioned in Section I, there are

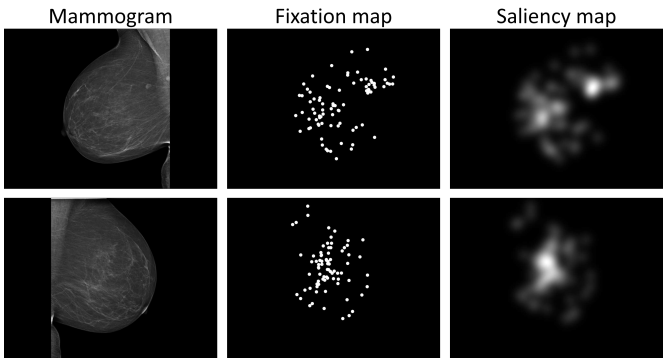


Fig. 1. Two examples of mammogram images and their corresponding binary fixation maps and saliency maps in the mammogram eye-tracking dataset. These two mammogram images are respectively the MLO view mammogram images of the right (top) and left (bottom) breasts from one anonymous case. The fixation maps show the fixation locations. The saliency maps are generated by giving rise to each fixation location a greyscale patch with a Gaussian distribution to simulate the size of the fovea.

a number of saliency models proposed in the literature. In this study, 10 traditional and 10 deep learning-based saliency models are selected, all of which have demonstrated state-of-the-art performance on the MIT300 [52], [53] (one of the most widely used natural image saliency benchmarks) and have made code of models publicly available by their authors. A brief introduction of the chosen saliency models can be found in Table I. We classify saliency models using deep neural networks as "deep learning-based models", and saliency models that mainly rely on extracting lower-level features (such as intensity, colour, and orientation) as "traditional models." In addition, according to previous studies [54], traditional models can be further categorised depending on the feature extraction mechanisms. For example, IttiKoch, GBVS, and Judd can be classified as cognitive models; GBVS belongs to graphical models; RARE2012, QSS, and signatureSal are spectral analysis models; Torralba and FES can be classified as Bayesian models; RARE2012 and Torralba are texture-based models; CovSal, FES, and LDS are statistics-based models. It should be noted that some models fall into more than one category. But they can be grouped into one general class, i.e., traditional models, as opposed to the approach of using deep neural networks.

C. Saliency evaluation metrics

The performance of saliency models is quantified by how well they can predict the ground truth gaze, rendered from human fixations via eye-tracking. To be able to quantitatively measure a saliency model's performance and compare different models, two forms of representation of the ground truth gaze are required: (1) fixation map; and (2) saliency map, as illustrated in Fig. 1.

A number of saliency evaluation metrics have been proposed [55]. There are six popular metrics, including Pearson's (or linear) correlation coefficient (CC), histogram intersection or similarity (SIM), Kullback-Leibler divergence (KLD), normalized scanpath saliency (NSS) and two variants of area under a curve (AUC) including Judd's AUC (AUC_J) and shuffled AUC (sAUC). These evaluation metrics capture different properties in saliency data, therefore produce different saliency model rankings. The study in [55] concluded that "specific tasks and applications may call for a different choice of metrics". Therefore, it is important to select appropriate evaluation metrics for a given application. In general, the above metrics can be classified into location-based and distribution-based metrics [56]. Location-based metrics evaluate "predicted" saliency maps by using "human" fixation maps as the ground truth, including NSS, AUC_J, and sAUC. Distribution-based metrics evaluate "predicted" saliency maps by using "human" saliency maps as the ground truth, including CC, SIM, and KLD. In our study, we target the regions of radiologists' visual interest and their relative importance, distribution-based metrics (i.e., CC, SIM, and KLD) are appropriate. In addition, the study in [55] suggests that, under normal assumptions, CC metric provides the fairest comparison between saliency models; SIM metric is a better fit when evaluating the relative importance of different image regions. In summary, CC and SIM metrics are most appropriate for our study and we therefore used them in our experiments.

A brief introduction to these evaluation metrics is given as follows, where $\mathbf{P}$ and $\mathbf{S}$ are the predicted and ground truth saliency maps respectively.

- CC is a statistical method to measure the correlation of two variables. It evaluates the accuracy of saliency prediction by:

$$\mathrm{CC}(\mathbf{P},\mathbf{S}) = \frac{\mathrm{cov}(\mathbf{P},\mathbf{S})}{\sigma(\mathbf{P})\cdot\sigma(\mathbf{S})}, \tag{1}$$

  where $\mathrm{cov}(\cdot)$ is the covariance and $\sigma(\cdot)$ is standard deviation. The value range of CC is between -1 and 1, and the closer the absolute value of CC is to 1, the more accurate is the predicted saliency.
- SIM measures the similarity between the predicted and the ground truth saliency maps by interpreting them as two 2-D histograms, $\mathbf{P}_i$ and $\mathbf{S}_i$. SIM can be calculated by:

$$\mathrm{SIM}(\mathbf{P},\mathbf{S}) = \sum_i \min(\mathbf{P}_i,\mathbf{S}_i), \tag{2}$$

  where i is the index of the horizontal axis of a histogram of the pixel intensity values; $\sum_i \mathbf{P}_i = \sum_i \mathbf{S}_i = 1$; and, the value range of SIM is between 0 and 1, and the higher the SIM value, the more similar are the two saliency maps so the more accurate is the predicted saliency.

D. Experiment setup

For model comparison, the parameters and settings of all traditional models were set to be the defaults provided by the original code. This group of traditional models is referred to as the "Traditional Group". In terms of deep learning-based models, all models were initialised with pre-trained weights (DVA is trained following the original publication' instructions, the weights of other models are provided by the original publications) on the SALICON dataset [63] but without any fine-tuning on the mammograms. This group of models is referred to as "DL Group". The SALICON is a

TABLE I
DETAILS OF 10 TRADITIONAL SALIENCY MODELS AND 10 DEEP LEARNING-BASED SALIENCY MODELS. TYPE T AND D REPRESENT TRADITIONAL MODELS AND DEEP LEARNING-BASED MODELS, RESPECTIVELY.

Model Name	Type	Backbone Network	Input Size (pre-training)	Advanced Features
IttiKoch [16]	T	–	1440×1080 px	Combining the colour, intensity, and orientation features at multiple spatial scales
GBVS [17]	T	–	1440×1080 px	Constructing Markov chain for feature maps extracted from a similar approach to *IttiKoch*
CovSal [18]	T	–	1440×1080 px	Calculating correlation of covariance matrices of simple local image features
FES [19]	T	–	1440×1080 px	Using sparse sampling and kernel density estimation to measure local feature contrast
LDS [20]	T	–	1440×1080 px	Learning discriminative subspaces to separate salient targets and distractors
RARE2012 [21]	T	–	1440×1080 px	Applying a multi-scale rarity mechanism on the colour and orientation features of the image
Judd [22]	T	–	1440×1080 px	Using linear support vector machine to use low-, mid-, and high-level image features
QSS [23]	T	–	1440×1080 px	Basing on the quaternion Fourier transform and uing eigenaxes and eigenangles
Torralba [24]	T	–	1440×1080 px	Combining bottom-up saliency, scene context, and top-down mechanisms in natural scenes
signatureSal [25]	T	–	1440×1080 px	Basing on sparse signal analysis to separate the foreground (target objects) and background
ML-Net [26]	D	VGG-16 [57]	640×480 px	Employing feature maps of three different layers and the prior map
DVA [27]	D	VGG-16	256×192 px	Using three sets of different-scale feature maps and decoder networks
SAM-VGG [28]	D	VGG-16	320×240 px	Simulating attention mechanism via a long short-term memory (LSTM)-based network
SAM-ResNet [28]	D	ResNet-50 [58]	320×240 px	Same as SAM-VGG
MSI-Net [32]	D	VGG-16	320×240 px	Using atrous spatial pyramid pooling (ASPP) [59] to fuse feature maps
EML-NET [29]	D	ResNet-50 (two-stream)	640×480 px	Using a two-stream feature extraction network pre-trained on ImageNet and PLACE separately
UNISAL [30]	D	MobileNetV2 [60]	384×288 px	Using multi-scale features by skip-connections
FastSal [33]	D	MobileNetV2	256×192 px	A computationally efficient model. The feature concatenation version is used here.
SalGAN [34]	D	VGG-16	256×192 px	Adopting a generative adversarial network (GAN) [61]
GazeGAN [31]	D	Modified U-Net	640×480 px	A GAN with U-Net [62] style

large-scale natural image saliency dataset that contains 10,000 training and 5,000 validation images with human attention behaviours recorded by mouse clicks instead of eye-tracking. The input sizes of these deep learning-based models are consistent with the authors' settings in Table I.

To explore the impact of transfer learning on mammograms' saliency prediction, all deep learning-based models were fine-tuned on mammograms based on the pre-trained weights on the SALICON dataset. This group of fine-tuned models is referred to as "DLFT Group". Note, this group of models can also be directly fine-tuned on mammograms without loading the pre-trained weights on the SALICON dataset so this gives us the opportunity to verify the necessity of pre-training on large-scale natural image saliency datasets. This verification would reveal whether saliency information learned from natural images would benefit learning saliency of medical images.

The eye-tracking mammogram dataset contains 196 mammograms from 98 cases. In order to obtain comprehensive and unbiased results of model performance, the dataset was divided into seven non-overlapping subsets and each subset contained 28 images from 14 cases, then k-fold Cross-Validation ($k = 7$ in our experiment) was applied to each deep learning-based model during fine-turning. For each fine-tuning and testing instance, one subset was kept as a test set, one as a validation set, and the remaining five subsets were used as a training set. The fine-tuning process for each instance was stopped and the best model was saved when the loss values on the corresponding validation set were consistently higher than the recorded minimum loss in five consecutive epochs. The report results were the mean performance of the best models of the seven tests. During the fine-tuning phase, the default loss function and optimizer were adopted for each model. According to the suggestions in the literature [26]–[34] and our empirical results (i.e., with the aim to obtain the lowest loss value in the validation set during fine-tuning), the hyper-parameters in each fine-tuning phase were set as follows:

1) The batch sizes for MSI-Net and GazeGAN were set to 1 due to the author's recommendation; for all other deep learning-based models were set to 4.
2) The initial learning rates for SAM-VGG, SAM-ResNet and GazeGAN were set to 2×10^{-4}; for ML-Net was set to 1×10^{-3}; for EML-NET was set to 3×10^{-3}; for MSI-Net was set to 1×10^{-5}; for UNISAL was set to 5×10^{-5}; for FastSal was set to 7×10^{-2}; for DVA was set to 2×10^{-4}.
3) For each model, the input image was resized by bilinear interpolation to the input size set by the authors for pre-training shown in Table I. In addition, as these models required input to be an RGB image format but the mammograms were grayscale images, we duplicated the grayscale values of an mammogram over the three RGB channels to generate the required input format to the models.
4) All models adopted default optimizers and learning rate adjustment strategies used in their original publications to ensure the fairness of the results.

In terms of statistical analysis, different statistical methods were used depending on the characteristics of the data. When the overall data from two independent samples was normally distributed (p-value$>$0.05 in Shapiro-Wilk tests), independent sample t-tests were used for statistical hypothesis testing; otherwise, Mann-Whitney U tests were used. Moreover, the Wilcoxon signed-rank tests were used for hypothesis testing of paired samples.

III. COMPARATIVE STUDY OF MODEL PROPERTIES

A. Investigation of different model genres

The performance of Traditional Group, DL Group, and DLFT Group on CC and SIM is shown in Fig. 2. It can be seen that the DL Group achieved higher scores than the Traditional Group in general. The models from DLTF Group are consistently highly ranked in both metrics. Some examples of the models' predictions are illustrated in Fig. 3. In order to

TABLE II
RESULTS OF THE MANN-WHITNEY U RANK TESTS FOR THE OVERALL PERFORMANCE OF THREE GROUPS ON METRICS CC AND SIM.

Group name	Group name	p-value (on CC)	p-value (on SIM)
Traditional	DL	$<$0.001	$<$0.001
Traditional	DLFT	$<$0.001	$<$0.001
DL	DLFT	$<$0.001	$<$0.001

investigate statistical differences between the performance of three groups, the Mann-Whitney U rank tests were performed on the results (i.e., Traditional Group versus DL Group, Traditional Group versus DLFT group, and DL Group versus DLFT group) on the metrics CC and SIM.

The statistical results are shown in Table II. The overall performance of the Traditional Group is significantly lower (p-value$<$0.001) than the DL group and the DLFT group. As can be visually assessed from Fig. 3, most models in the Traditional Group have inadequate performance in predicting the saliency of mammograms. One plausible reason for the low performance is that features in these models are specifically designed for natural images and are not specific to mammography. For example, Judd and Torralba compute image saliency by detecting features such as faces or natural objects, and it may not be feasible to extrapolate this to mammograms. Besides, it is worth noting that LDS is the best performing traditional model on the mammogram dataset, which outperforms most models in the DL Group. This suggests that the saliency detection method adopted by LDS, i.e., defining discriminative subspaces to separate salient targets and distractors, might be beneficial for mammogram saliency detection.

With regard to deep learning-based models, models in the DL Group are trained on the large-scale natural image saliency dataset, i.e., SALICON, then applied to mammograms directly. It can be seen from the examples in Fig. 3, the agreement between the output of the DL Group's models and ground truth is still insufficient, although their overall performance is better than the Traditional Group. The results lead to a similar conclusion where saliency features for natural images are not fully suitable for mammograms. Deep learning-based models can be fine-tuned on the mammogram dataset by supervised learning to obtain mammograms' saliency-related features. The results in Table II indicate that the overall performance of fine-tuned deep learning models, i.e., the DLFT Group, substantially outperforms (p-value$<$0.001) the other two groups. The saliency maps of the DLFT Group are more consistent with the ground truth as illustrated in Fig. 3. There is evidence to show that using appropriate deep learning-based models and training them on mammogram saliency data achieve good performance for saliency prediction on mammograms.

To have an insightful view on the usage of these models for medical imaging, it is worthwhile to investigate the correlation in performance rankings between mammograms and natural images. To this end, we calculated the Kendall rank correlation coefficient (KRCC) between the performance of these models on natural images and mammograms. Note, the data of natural images is based on the MIT/Tuebingen Saliency Benchmark [53]. To make a fair analysis, Torralba and FastSal were excluded because they were not present in the MIT/Tuebingen benchmark, and EML-NET was excluded because its model version published on the MIT/Tuebingen benchmark was inconsistent with the implementation in the original publication. The comparison of rankings is shown in Table III and Table IV. The Kendall rank correlation coefficient (KRCC) between the performance on natural images and mammograms for the Traditional Group, DL Group, and DLFT Group on CC are 0.2778, 0.2143, and -0.2143 respectively; and on SIM are 0.3889, 0.3571, and 0.2857 respectively. A weak correlation is observed between these groups, i.e., the correlation in model performance rankings between mammograms and natural images does not exceed 0.5. This indicates that the performance of saliency models on these two domains is not consistent, and that models that perform well on natural images are not necessarily good models for medical imaging, and vice versa. This also implies that the selection of saliency models for medical imaging cannot rely on the existing saliency benchmarks for natural images. To provide a reliable reference for model selection, existing models should be implemented and tested using medical images.

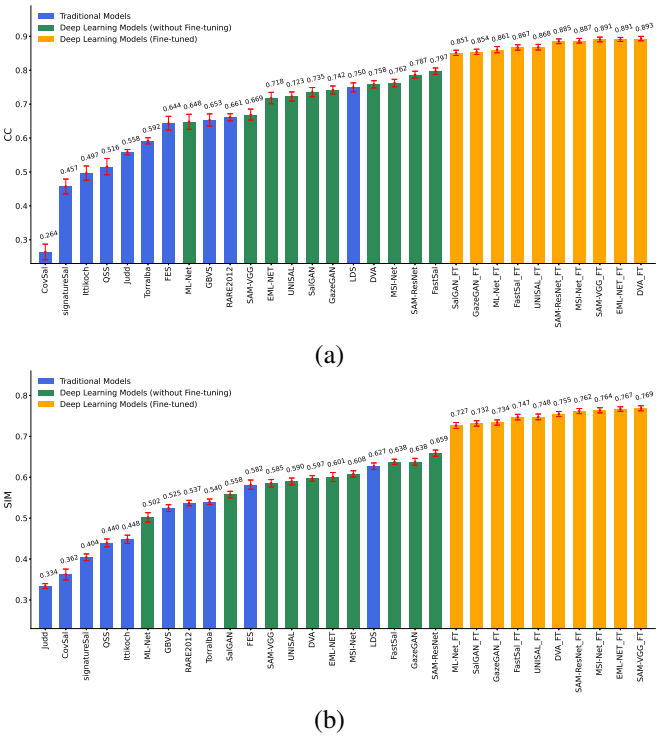


Fig. 2. The results of all saliency models, including 10 traditional models (blue bars), 10 deep learning-based models (green bars), and 10 fine-tuned (on mammograms) deep learning-based models (orange bars) on CC and SIM are demonstrated in (a) and (b), respectively. The error bar represents a 95% confidence interval.

B. Investigation of different deep learning network structures

To advance deep learning-based saliency modelling for medical imaging, it is vital to explore what kinds of network structures are suitable for this specific application. Now, we investigate the impact different network structures have on

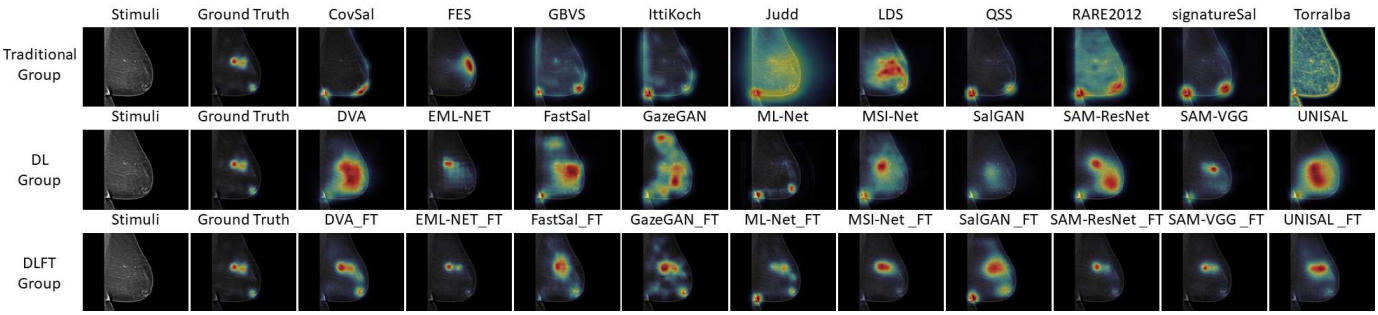


Fig. 3. Prediction results of all models for the same mammogram's saliency. The first and second columns on the left show the mammogram image (Stimuli) and the corresponding saliency map (Ground Truth) respectively. The other heatmaps are the predictions of saliency models. The heatmaps in the first row is from the traditional Group, the second row is from the DL Group, and the third row is from the DLFT Group.

TABLE III
RANKING OF THE PERFORMANCE OF TRADITIONAL SALIENCY MODELS ON THE EYE-TRACKING MAMMOGRAM DATASET AND ON NATURAL IMAGES (I.E., MIT300 DATASET).

	CC	SIM
Model name	mammograms / natural images	mammograms / natural images
LDS	1st / 1st	1st / 1st
RARE2012	2nd / 6th	3rd / 5th
GBVS	3rd / 4th	4th / 4th
FES	4th / 3rd	2nd / 3rd
Judd	5th / 5th	9th / 8th
QSS	6th / 8th	6th / 7th
Ittikoch	7th / 9th	5th / 9th
signatureSal	8th / 7th	7th / 6th
CovSal	9th / 2nd	8th / 2nd

TABLE IV
RANKING OF THE PERFORMANCE OF FINE-TUNED AND NOT FINE-TUNED DEEP LEARNING-BASED SALIENCY MODELS ON THE EYE-TRACKING MAMMOGRAM DATASET AND ON NATURAL IMAGES (I.E., MIT300 DATASET).

Model name	CC (fine tuned / not fine tuned on mammograms / natural images)	SIM (fine tuned / not fine tuned on mammograms / natural images)
DVA	1st / 3rd / 7th	4th / 4th / 7th
SAM-VGG	2nd / 7th / 8th	1st / 6th/ 5th
MSI-Net	3rd / 2nd / 2nd	2nd / 3rd / 2nd
SAM-ResNet	4th / 1st / 4th	3rd / 1st / 4th
UNISAL	5th / 6th / 1st	5th / 5th / 1st
ML-Net	6th / 8th / 6th	8th / 8th / 8th
GazeGAN	7th / 4th / 3rd	6th / 2nd / 3rd
SalGAN	8th / 5th / 5th	7th / 7th / 6th

TABLE V
THE MEAN PERFORMANCE OF FINE-TUNED MODELS USING DIFFERENT BACKBONES ON CC AND SIM.

Backbone name	CC↑	SIM↑
MobileNetV2	0.8676	0.7473
VGG-16	0.8767	0.7490
ResNet-50	0.8882	0.7643

TABLE VI
RESULTS OF INDEPENDENT SAMPLES t-TESTS FOR FINE-TUNED MODELS USING DIFFERENT BACKBONES.

Backbone name	Backbone name	p-value (on CC)	p-value (on SIM)
VGG-16	ResNet-50	$<$0.001	$<$0.001
VGG-16	MobileNetV2	$<$0.010	$>$0.050
ResNet-50	MobileNetV2	$<$0.001	$<$0.001

TABLE VII
RESULTS OF THE MANN-WHITNEY U RANK TESTS FOR FINE-TUNED MODELS USING ONE-STREAM OR TWO-STREAM BACKBONES.

Metrics	Means (one-stream)	EML-NET (two-stream)	p-value
CC↑	0.8732	0.8889	$<$0.001
SIM↑	0.7484	0.7652	$<$0.001

TABLE VIII
THE RESULTS OF THE WILCOXON SIGNED-RANK TESTS ON FINE-TUNED EML-NET USING DIFFERENT BACKBONES

Metrics	One-stream (ImageNet / PLACE)	Two-stream	p-value
CC↑	0.8860 / 0.8832	0.8909	$<$0.010 / $<$0.010
SIM↑	0.7605 / 0.7579	0.7668	$<$0.010 / $<$0.010

performance. Statistical significance tests were performed on the fine-tuned models adopting different architectures. From the perspective of backbone networks, there are three main types involved: VGG-16 (used by ML-Net, DVA, SAM-VGG, MSI-Net, and SalGAN), ResNet-50 (used by SAM-ResNet and EML-NET), and MobileNetV2 (used by UNISAL and FastSal). The mean performance of models using different backbones on CC and SIM is shown in Table V, and the results of t-tests for comparing model performance are shown in Table VI. Unlike other models using a one-stream backbone, EML-NET uses a two-stream backbone as the feature extractor. The Mann-Whitney U rank tests were performed on the performance of models using one-stream and two-stream backbones and the results are shown in Table VII. Furthermore, the authors of EML-NET also provide one-stream version models for investigating the influence of one- and two-stream backbone on saliency prediction, and their performance on the mammograms is shown in Table VIII. In addition, contrary to other models, there are two models (i.e., GazeGAN and SalGAN) based on generative adversarial network (GAN). The mean CC and SIM for models with GAN is 0.8528 and 0.7327, and for models without GAN is 0.8802 and 0.7544. The differences in means between CCs and between SIMs were statistically significant (independent samples t-tests, p-value$<$0.001).

Backbone networks are important in extracting image features for saliency detection. The backbone networks in this study include ResNet-50, VGG-16, and MobileNetV2, of which ResNet-50 has the largest parameter scale, followed by

VGG-16, and the parameter scale of MobileNetV2 is much smaller than ResNet-50 and VGG-16. It is by now generally accepted that when performing complex tasks with sufficient computational resources and data volume, deep neural networks with large-scale parameters, such as ResNet-50 and VGG-16, perform better than lightweight neural networks. But the backbones with fewer trainable parameters, such as MobileNetV2, can significantly save computational resources at the cost of performance accuracy. As can be seen in Table V, the models applying ResNet-50 as a backbone have significantly higher performance (p-value$<$0.001) than models using VGG-16 and MobileNetV2 as their backbones. Besides, the mean performance of models using VGG-16 as backbones is higher than models using MobileNetV2, and there is a significant difference (p-value$<$0.01) in the CC metric. These results are in line with the general observations mentioned above, which implies that complex deep neural networks with strong representation ability are more suitable for predicting mammograms' visual saliency. The only model that ranks in the top two positions on both evaluation metrics is EML-NET, and its main structural difference from other models is that it uses a two-stream backbone to extract features. This two-stream backbone is composed of two ResNet-50s in parallel, which are loaded with parameters pre-trained on ImageNet and PLACE datasets respectively. We trained this model on SALICON, then fine-tuned on the mammogram dataset in our study. We found that a two-stream backbone could acquire more prior knowledge than a one-stream backbone to improve the predictive power of the model, but with a increase in the training difficulty and computational resource consumption of the network. According to Table VII and Table VIII, the two-stream backbone version of EML-NET is not only better than the one-stream backbone version of its own, but also significantly outperforms other models with a single-stream backbone (p-value$<$0.001). This implies that the use of broader or multimodal prior knowledge is beneficial for mammograms' saliency prediction.

Although GANs currently perform well in many image generation tasks, models using GAN have no benefit for the mammograms' saliency prediction task. According to the experimental results, GAN style models have significant lower performance than the deep learning models without the use of GAN. This may be due to the fact that GANs are more difficult to train than other networks. Therefore, GANs should be used with caution if large-scale datasets are not available in the application domain.

C. Investigation of transfer learning methods

Transfer learning is a common approach to applying deep learning algorithms to medical image related tasks, which mainly consists of pre-training on datasets of unrelated tasks and fine-tuning on target data. Collecting a large-scale saliency dataset for medical images is nontrivial due to the limited access to radiologists in practice; and so far, the publicly available medical saliency datasets are too small to train a saliency model for medical images. Therefore, our intention is to investigate whether and to what extent the larger saliency dataset available for natural images could be leveraged to predict saliency of medical images. This would provide practical solutions for medical applications. To verify the hypothesis that saliency data available for natural images is useful for learning saliency of medical images, we compare deep learning-based models fine-tuned with and without pre-trained weights on the natural image saliency SALICON dataset.

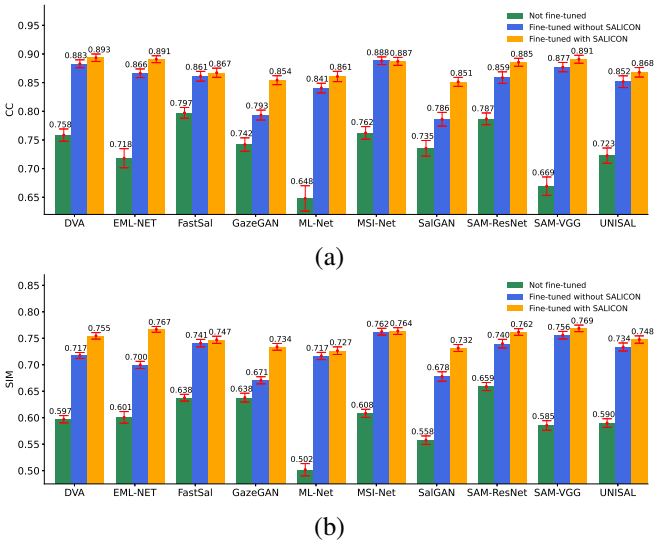


Fig. 4. Performance comparison of deep learning-based models that were not fine-tuned, fine-tuned with and without pre-trained weights on the SALICON dataset, using metrics CC (a) and SIM (b). The green, blue and orange bars represent models that were not fine-tuned, fine-tuned with and without the SALICON dataset, respectively. The error bar represents a 95% confidence interval.

TABLE IX
THE MEAN PERFORMANCE OF MODELS USING DIFFERENT FINE-TUNING STRATEGIES ON CC AND SIM.

Models	CC↑	SIM↑
Not fine-tuned	0.7340	0.5975
Fine-tuned without SALICON	0.8506	0.7215
Fine-tuned with SALICON	0.8749	0.7502

The results are shown in Fig. 4, where the models contained in the DLFT Group were fine-tuned with and without the SALICON dataset. The details of the three different training strategies involved in Fig. 4 and Table IX are as follows:

- Not fine-tuned: The networks were initialised with the model pre-trained on SALICON, then directly tested on the eye-tracking mammogram dataset without any fine-tuning.
- Fine-tuned without SALICON: The backbone networks were first initialised with parameters pre-trained on large natural image datasets such as ImageNet and PLACE (except for GazeGAN, which was initialised from a Gaussian distribution by its authors [31]), and the downstream networks were initialised by the default method provided by the authors. Then the network was fine-tuned on the eye-tracking mammogram dataset.
- Fine-tuned with SALICON: The networks were first initialised with the model pre-trained on SALICON and then fine-tuned on the eye-tracking mammogram dataset.

The mean scores of CC and SIM are shown in Table IX.

The results of the Wilcoxon signed-rank tests indicates that models fine-tuned with saliency information of natural images can significantly ($p < 0.001$) improve their performance for predicting saliency of medical images.

We found that saliency information learned from a large-scale natural image dataset, i.e., the SALICON dataset, is beneficial for learning saliency of mammograms. According to Fig. 4, in general, fine-tuned models with the pre-training weights in SALICON give better performance than those without SALICON. In terms of the mean performance presented in Table IX and the Wilcoxon signed-rank tests, the models fine-tuned on the basis of the parameters pre-trained in SALICON significantly outperform the models that are not fine-tuned with SALICON. These results can be attributed to the following reasons. First, models can obtain saliency-related features from the SALICON dataset that are otherwise difficult to learn in a smaller-scale saliency dataset. The second reason is that loading the pre-trained weights on SALICON can initialise all parameters in the model, which allows the networks to be more stable and enhance their convergence. When SALICON weights are not loaded, the default weight initialisation method of the considered deep learning-based models is to load the weights trained based on large-scale natural image data sets (such as ImageNet and PLACE datasets) for their backbone networks (except for GazeGAN, which is initialised from a Gaussian distribution [31]). This initialisation method does not initialise the parameters of their downstream networks, such as decoder networks. This default initialisation method makes it more difficult to obtain optimal performance by directly fine-tuning on the mammogram dataset. Based on above analyses, pre-training models on larger-scale natural image saliency datasets (e.g., SALICON) before fine-tuning them on a limited-scale mammogram dataset is suggested. It is worth noting that only MSI-Net is barely affected by SALICON fine-tuning. Amongst all models, MSI-Net is the only model with its encoder trained on both ImageNet and PLACE dataset; other models' encoder was trained on one dataset only (i.e., either ImageNet or PLACE). Both ImageNet and PLACE are visual recognition datasets that are not directly relevant for saliency. This implies that if pre-training on one visual task is sufficient, the model is readily transferred to learn the new task of predicting saliency of mammograms. This may explain the fact that MSI-Net does not significantly benefit from further training on SALICON and is readily equipped with sufficient capacity for fine-tuning on mammograms. In addition, based on the results in Table IX and the results of the Wilcoxon Signed-Rank tests, fine-tuning a model rather than directly applying any existing saliency models on the mammogram dataset significantly improves the model performance (p-value$<$0.001).

D. Significant findings

The above systematic experiments and statistical analyses have revealed the following findings:

- The fine-tuned deep learning-based saliency models represent one of the most effective type of methods for predicting radiologists' visual attention in mammography diagnosis.
- The saliency benchmarks or model rankings developed for natural images cannot provide a reliable reference for the selection of models for medical image saliency prediction.
- Complex deep neural networks with strong representation ability, e.g., ResNet-50 and two-stream backbone, are more suitable for predicting mammograms' visual saliency.
- The GANs should be used with caution if large-scale datasets are not available in the application domain.
- The ability of deep learning-based saliency models to predict radiologists' gaze in mammogram reading can be significantly improved by transfer learning. In particular, model performance can be improved by pre-training the model on a large-scale natural image saliency dataset and then fine-tuning it on the target medical image dataset.

IV. Proposed Parallel Multi-Stream Encoded Model for Saliency Prediction of Mammograms

Based on our findings in Section III-D, we hereby propose a parallel multi-stream model, using an encoder-decoder architecture, as depicted in Fig. 5. The encoder integrates three state-of-the-art backbone networks in parallel that considers introducing stronger and more extended image representations for mammogram saliency prediction, and including diversified features from CNN- and transformer-based networks and deep encoded high-resolution representations.

A. Model details

1) Encoding: In order to obtain strong and diversified image representations, three parallel state-of-the-art backbone networks, including *ConvNeXt* (version of ConvNeXt-B) [64], *HRNet* (HRNet-W48-C) [65], and *CSwin Transformer* (CSwin-B) [66], are chosen to construct an image encoder. In selecting these backbone models, the intention is to use well-performing models with proven efficacy but representing distinctive modelling philosophies. ConvNeXt is a representative pure CNN-based backbone with a classical architecture. It contains advancements from well-established backbone networks; and it provides low-resolution representations in the deep layers and high-resolution representations in the shallow layers. HRNet connects high-to-low resolution convolution streams in parallel as opposed to scaling down the size of the feature map in steps (the approach taken by most classical backbones like ConvNeXt), which results in spatially more precise learned representations. CSwin Transformer adopts a new mechanism for computing self-attention in the horizontal and vertical stripes in parallel that form a cross-shaped window. It can provide sufficient long-range information that is lacking in ConvNeXt and HRNet. For each backbone, four sets of feature maps with the resolutions of $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{16}$, and $\frac{1}{32}$, respectively are extracted for the decoder. For both ConvNext and CSwin Transformer, these feature maps are extracted from the end of their four feature resolution stages. For HRNet, they are extracted from the final stage. In addition, the spatial size of the input images of ConvNeXt and HRNet is 384×288, while the images fed into CSwin Transformer are first zero-padded to a spatial size of 384×384 to match its input requirement and then cropped after backbone processing to align with the spatial size of feature maps from other backbones.

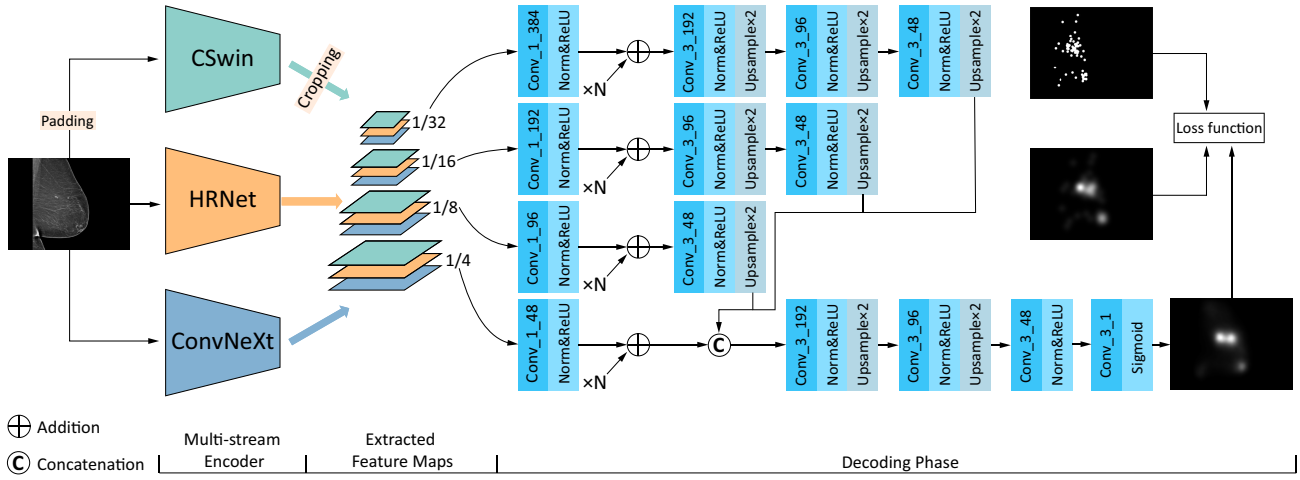


Fig. 5. The architecture of proposed parallel multi-stream encoded model. The left-hand side shows the encoders, and the right-hand side shows the details of the decoder. The different spatial sizes of feature maps are extracted by parallel backbones (CSwin Transformer, HRNet, and ConvNeXt, from top to bottom) and feed into the decoder. *N* equals the number of encoders employed; *Conv_m_n* represents convolutional layer with $m \times m$ kernel and n output channels.

2) Decoding: To eliminate the potential influence of different channels of backbones' outputs on the results, the outputs from these backbones are processed by 1×1 convolution layers to align the number of output channels. The feature maps with the same resolution are fused by element-wise addition. Following this, those with a resolution of less than $\frac{1}{4}$ from the encoders are separately concatenated and processed by a series of convolution and upsampling operations (i.e., nearest-neighbor interpolation) to restore the spatial dimension of the feature maps to $\frac{1}{4}$ resolution. Subsequently, the restored feature maps and the maps with a resolution of $\frac{1}{4}$ are concatenated and processed by a series of convolution and upsampling operations to restore the spatial dimension to the same size of the input image and reduce the channel dimensions to generate the final saliency map.

3) Loss function: Using the saliency evaluation metrics to define the loss function has achieved notable success in saliency prediction [28]–[31]. Accordingly, we adopted a linear combination of four metrics as the loss function to train our model, including the Normalized Scanpath Saliency (NSS), Kullback-Leibler divergence (KLD), Linear Correlation Coefficient (CC), and Similarity (SIM). Let $\mathbf{y}^s$, $\mathbf{y}^f$, and $\hat{\mathbf{y}}$ be the ground truth saliency map, fixation map, and predicted saliency map, our loss function is defined as:

$$\begin{aligned}\mathcal{L}(\mathbf{y}^s,\mathbf{y}^f,\hat{\mathbf{y}}) =& \lambda_1\mathcal{L}_{\text{NSS}}(\mathbf{y}^f,\hat{\mathbf{y}}) + \lambda_2\mathcal{L}_{\text{KLD}}(\mathbf{y}^s,\hat{\mathbf{y}}) \\ &+ \lambda_3\mathcal{L}_{\text{CC}}(\mathbf{y}^s,\hat{\mathbf{y}}) + \lambda_4\mathcal{L}_{\text{SIM}}(\mathbf{y}^s,\hat{\mathbf{y}}),\end{aligned} \tag{3}$$

where λ_1, λ_2, λ_3, and λ_4 are the weights of individual metrics. More details of the loss function can be found at [67].

B. Experiments and results

1) Training strategies: According to the analysis in section III-C, pre-training the model on SALICON dataset is beneficial for predicting the visual attention of radiologists. Therefore, the models were first pre-trained on SALICON, and then fine-tuned and validated by the same k-fold Cross-Validation ($k = 7$) strategy as detailed in section II-D on mammogram dataset. The pre-training process contains two stages: encoder training and decoder training. In the stage of encoder training, each selected backbone is used as an encoder to form an encoder-decoder network with a similar architecture shown in Fig. 5 and trained on SALICON. Then all the trained encoders form a parallel encoder. In the stage of deocder training, the target network consists of a trained parallel encoder and an untrained decoder, where the parameters of the trained parallel encoder are all fixed, and then only the decoder is trained on SALICON. The fine-tuning process is similar to the pre-training process. The models with a single backbone encoder are first loaded the parameters obtained from the pre-training phase and fine-tuned on mammogram dataset. Then the parallel encoder and decoder of the target network are loaded the parameters fine-tuned on mammogram dataset and the parameters pre-trained on SALICON respectively, where the parameters of the parallel encoder are all fixed, and then only the decoder is fine-tuned on mammogram dataset. Because the models are pre-trained on SALICON dataset, the λ_1, λ_2, λ_3, and λ_4 in loss function are set to -1, 10, -2, and -1 respectively according the previous study [67]. The learning rate is set to 2×10^{-5} and 1×10^{-4} for pre-training phase and fine-tuning phase respectively, which is then multiplied by 0.1 for every 2 epochs. Models are trained with a batch size of 4 for 30 epochs with a stop patience of 5 epochs.

2) Results: The results are listed in Table X. For all three backbones, using two backbones to form a parallel encoder (two-stream encoder) is more beneficial than using a single backbone as an encoder for mammogram saliency predictions. In addition, the model using all three backbones to construct a parallel encoder (a three-stream encoder) further improves the performance compared to models using a two-stream encoder. By introducing parallel backbones, the model achieves the best performance scores on CC and SIM in all saliency models.

To further validate the proposed method, the Wilcoxon Signed-Rank tests are performed on the seven model variants in Table X show the following results. 1) for the three variants that adopt a single encoder, there is no statistically

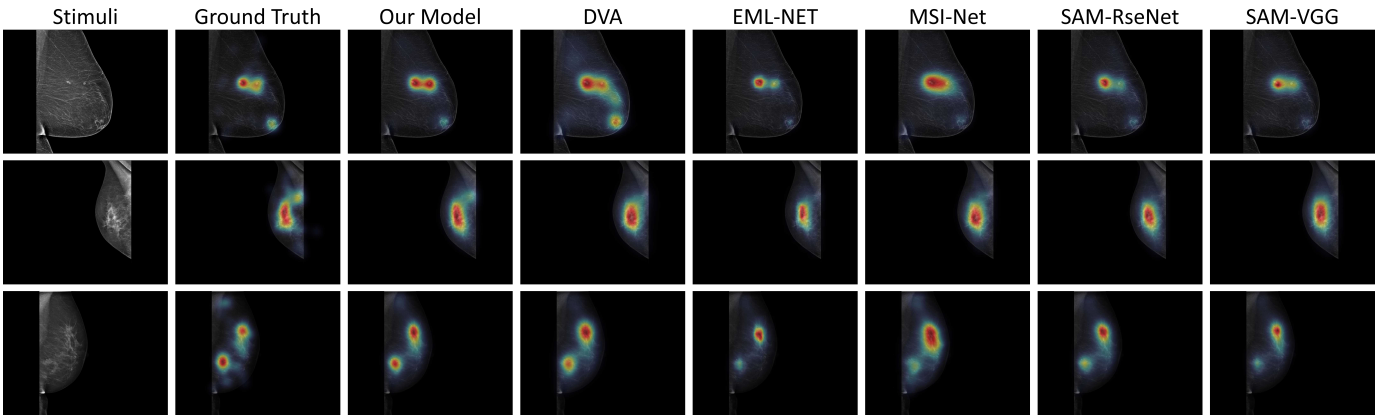


Fig. 6. Comparison of saliency maps generated by our model and the state-of-the-art models in the comparative study (see Section III). From left to right, the columns correspond to mammogram images (Stimuli), the corresponding saliency maps (Ground Truth), and the prediction results from our model and the state-of-the-art saliency models, respectively.

TABLE X
EXPERIMENT RESULTS: MAMMOGRAM SALIENCY PREDICTION OF SINGLE AND PARALLEL MULTI-STREAM ENCODED MODELS WITH DIFFERENT BACKBONES. μ AND σ REPRESENT THE MEAN AND STANDARD DEVIATION OF k-FOLD CROSS-VALIDATION (k=7), RESPECTIVELY. **BOLD** FONT INDICATES THE BEST PERFORMANCE

ConvNeXt	HRNet	CSwin Transformer		CC↑	SIM↑
✓	–	–	μ	0.9007	0.7775
			σ	0.0061	0.0090
–	✓	–	μ	0.9017	0.7785
			σ	0.0070	0.0094
–	–	✓	μ	0.8995	0.7762
			σ	0.0046	0.0085
✓	✓	–	μ	0.9043	0.7813
			σ	0.0077	0.0098
✓	–	✓	μ	0.9030	0.7802
			σ	0.0060	0.0097
–	✓	✓	μ	0.9038	0.7809
			σ	0.0063	0.0095
✓	✓	✓	μ	**0.9061**	**0.7830**
			σ	0.0070	0.0096

TABLE XI
COMPARISON OF THE PERFORMANCE OUR MODEL AND THE STATE-OF-THE-ART MODELS IN THE COMPARATIVE STUDY (SEE SECTION III) ON THE MAMMOGRAM DATASET. μ AND σ REPRESENTS THE MEAN AND STANDARD DEVIATION OF k-FOLD CROSS-VALIDATION (k=7), RESPECTIVELY. **BOLD** FONT INDICATES THE BEST PERFORMANCE

Model Name		CC↑	SIM↑	NSS↑	AUC_J↑
SAM-ResNet [28]	μ	0.8855	0.7618	2.9095	0.9417
	σ	0.0100	0.0117	0.0447	0.0033
MSI-Net [32]	μ	0.8871	0.7636	2.8867	0.9418
	σ	0.0125	0.0152	0.0432	0.0038
SAM-VGG [28]	μ	0.8908	0.7687	2.9503	0.9426
	σ	0.0090	0.0123	0.0257	0.0039
EML-NET [29]	μ	0.8909	0.7668	2.9876	0.9435
	σ	0.0064	0.0098	0.0396	0.0030
DVA [27]	μ	0.8935	0.7546	2.9212	0.9425
	σ	0.0086	0.0138	0.0450	0.0036
Our Model	μ	**0.9061**	**0.7830**	**3.0109**	**0.9446**
	σ	0.0070	0.0096	0.0337	0.0032

significant difference (i.e., p-value$>$0.05) in performance; 2) for the three variants that adopt a two-stream encoder, there is no statistically significant difference (i.e., p-value$>$0.05) in performance; 3) models with a two-stream encoder have significantly higher (i.e., p-value$<$0.01) performance than the models with a single encoder (note that is used to form the two-stream encoder); 4) the three-stream variant significantly outperforms (i.e., p-value$<$0.01) all models that adopt either a two-stream encoder or a single encoder. These results show that the proposed method can significantly improve the performance of predicting mammograms' saliency.

3) Comparison with state-of-the-art methods: The state-of-the-art visual saliency models in Section III are selected (Top-ranked models in Fig. 2) for the performance comparison on the mammogram dataset. The overall performance results are reported in Table XI. It can be seen that our model achieves the best performance on CC and SIM. In addition, two widely used location-based saliency evaluation metrics, namely NSS and AUC_J are used, as the complement to the distribution-based metrics. It can be seen that our model achieves the best performance on NSS and AUC_J as well. Fig. 6 shows saliency maps generated by our models and other models for mammograms. By visually assessing these saliency maps, it can be seen that our model is in closer agreement with the ground truth than other models.

V. DISCUSSION

In many clinical applications, a computational model's prediction accuracy is rather critical for its successful deployment in real practice. The focus of the proposed model in Section IV is to achieve a most accurate possible saliency prediction by using strong and diversified image representations. The increased accuracy is often achieved at the expense of computational costs. It is worth discussing the model's complexity in terms of the number of parameters as well as runtime (indicators of the computational resources required for a deep learning-based model's inference). This will also direct the future research in managing the trade-off between accuracy and complexity for specific clinical environments. A

TABLE XII
THE NUMBER OF PARAMETERS AND RUNTIME FOR SOTA MODELS IMPLEMENTED IN THIS STUDY AND THE PROPOSED MODEL. RUNTIME INDICATES THE RUNTIME OF A MODEL TO PREDICT SALIENCY FOR A SIGNLE MAMMOGRAM IMAGE.

Model name	Size (M)	Runtime (s)
DVA	20.60	0.05
MSI-Net	24.93	0.05
EML-NET	47.09	0.04
SAM-VGG	51.84	0.06
SAM-ResNet	70.09	0.08
Our model	240.36	0.13

comparison of the number of parameters and runtime (running on a single NIVIDA GTX 1080 GPU) of the proposed model with stat-of-the-art saliency models is illustrated in Table XII. Our proposed model has a larger number of parameters and a longer runtime compared to other models since it combines multiple complex deep networks to form its encoder. Again, in this study we have concentrated on developing a most accurate model for use in clinical applications; and the proposed model is a proof of concept, harvesting the significant findings of the comparative study in Section. III, *"complex deep neural networks with strong representation ability are more suitable for predicting mammograms' visual saliency."* To this end, we have taken the approach of combing multiple deep networks to produce a plausible solution. Although achieving high accuracy is a top priority than saving computational resources for many clinical applications, it should be noted that a trade-off between the need for computing power, inference time, and accuracy is often considered in practical settings. The future research includes further optimisation of the proposed model to reduce the required computing capacity while maintaining its accuracy.

For our proposed saliency model, the weights of the four subloss functions are determined empirically. As per previous visual saliency prediction studies [28], [31], [67], the actual tuning process consists of two main steps. In the first step, a set of initial weights of sublosses is produced with the aim to make the impact of individual subloss functions on the overall model result relatively consistent/equal. In the second step, a "grid search"-like method, i.e., fine-tuning individual weights by adjusting one weight and fixing the remaining weights to optimise the performance on the validation set of SALICON [31]. The goal is to find a combination of weights that allows the model to achieve good and balanced scores on the commonly used saliency evaluation metrics. More complex weights determination methods e.g., automatic hyperparameter optimisation may yield better results for specific application scenarios. However, this is beyond the scope of this study, and is worth further exploring in future research.

In addition to predicting the spatial fixation distribution/density as discussed in this paper, research has been undertaken to predict human scanpaths in natural images [68]–[72] and achieved promising results. This demonstrates the potential for the development of scanpaths prediction models for medical images, which is highly relevant for improving radiologists' diagnostic performance.

It should be noted that deep learning models are data-driven, and the scale of datasets affects their performance [73]. With respect to medical imaging-related tasks, recent studies have demonstrated that the performance of deep learning algorithms can be improved by increasing the size of the training dataset [74]. To the best of our knowledge, the mammogram saliency dataset used in this study is the largest of its kind but remains limited. We would expect that increasing the size of the dataset would benefit the prediction power of deep learning-based models.

VI. CONCLUSION

In this paper, we have investigated 20 state-of-the-art saliency models in predicting the visual attention of radiologists during the reading of mammograms. We found that the deep learning-based models developed for natural images are effective to an extent after appropriate fine-tuning using eye-tracking data of mammograms. However, the saliency benchmarks or model rankings resulted from natural images cannot provide a reliable reference for the selection of models for medical imaging. In addition, we found that complex deep neural networks with strong representation ability and pre-training models on large-scale natural saliency datasets (i.e., SALICON) are beneficial for predicting the visual saliency of mammograms. Following this, we have developed a saliency model that uses parallel multi-stream encoders to predict the saliency of the mammogram, which achieves superior performance to the existing state-of-the-art models.

Despite saliency models have been applied to improve clinical diagnosis and computer-aided diagnostic systems [75], this work provides previously unavailable analyses and guidelines to inform the design of better saliency models for diagnostic imaging. Future work includes integrating our method and model to real-world clinical settings and applications.

REFERENCES

[1] E. E. M. Stewart, M. Valsecchi, and A. C. Schütz, "A review of interactions between peripheral and foveal vision," *J. Vis.*, vol. 20, no. 12, p. 2, 11 2020.
[2] R. Bertram *et al.*, "Eye Movements of Radiologists Reflect Expertise in CT Study Interpretation: A Potential Tool to Measure Resident Development," *Radiol.*, vol. 281, no. 3, pp. 805–815, 2016.
[3] G. Tourassi *et al.*, "Investigating the link between radiologists' gaze, diagnostic decision, and image content," *J. Amer. Med. Inform. Assoc.*, vol. 20, no. 6, pp. 1067–1075, 06 2013.
[4] Y. Li *et al.*, "Computational modeling of human reasoning processes for interpretable visual knowledge: a case study with radiographers," *Sci. Rep.*, vol. 10, no. 1, p. 21620, 12 2020.
[5] H. Ashraf *et al.*, "Eye-tracking technology in medical education: A systematic review," *Med. Teacher*, vol. 40, no. 1, pp. 62–69, 2018.
[6] N. Khosravan *et al.*, "A collaborative computer aided diagnosis (C-CAD) system with eye-tracking, sparse attentional model, and deep learning," *Med. Image Anal.*, vol. 51, pp. 101–115, 2019.
[7] S. Banerjee, S. Mitra, and B. Uma Shankar, "Automated 3D Segmentation of Brain Tumor Using Visual Saliency," *Inf. Sci.*, vol. 424, no. C, p. 337–353, Jan. 2018.
[8] J. Bernal *et al.*, "WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Comput. Medical Imag. Graph.*, vol. 43, p. 99—111, July 2015.
[9] Y. Yuan, D. Li, and M. Q.-H. Meng, "Automatic Polyp Detection via a Novel Unified Bottom-Up and Top-Down Saliency Approach," *IEEE J. Biomed. Health Inform.*, vol. 22, no. 4, pp. 1250–1260, 2018.
[10] N. Stec *et al.*, "A systematic review of fatigue in radiology: Is it a problem?" *Am. J. Roentgenol.*, vol. 210, no. 4, pp. 799–806, 2018.

[11] I. Pershin *et al.*, "Artificial intelligence for the analysis of workload-related changes in radiologists' gaze patterns," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 9, pp. 4541–4550, 2022.

[12] H. Ashraf *et al.*, "Eye-tracking technology in medical education: A systematic review," *Med. Teach.*, vol. 40, no. 1, pp. 62–69, 2018.

[13] J. Born *et al.*, "On the role of artificial intelligence in medical imaging of covid-19," *Patterns*, vol. 2, no. 6, p. 100269, 2021.

[14] A. Tariq *et al.*, "Current clinical applications of artificial intelligence in radiology and their best supporting evidence," *J. Am. Coll. Radiol.*, vol. 17, no. 11, pp. 1371–1381, 2020.

[15] H. Chen *et al.*, "Explainable medical imaging ai needs human-centered design: Guidelines and evidence from a systematic review," *npj Digit. Med.*, vol. 5, no. 1, 2022.

[16] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural Netw.*, vol. 19, no. 9, pp. 1395–1407, 2006, brain and Attention.

[17] J. Harel, C. Koch, and P. Perona, "Graph-Based Visual Saliency," in *Proc. 19th Int. Conf. Neural Inf. Process. Syst.*, ser. NIPS'06. Cambridge, MA, USA: MIT Press, 2006, p. 545–552.

[18] E. Erdem and A. Erdem, "Visual saliency estimation by nonlinearly integrating features using region covariances," *J. Vis.*, vol. 13, no. 4, p. 11, 03 2013.

[19] H. Rezazadegan Tavakoli, E. Rahtu, and J. Heikkilä, "Fast and Efficient Saliency Detection Using Sparse Sampling and Kernel Density Estimation," in *Image Anal.*, A. Heyden and F. Kahl, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 666–675.

[20] S. Fang *et al.*, "Learning Discriminative Subspaces on Random Contrasts for Image Saliency Analysis," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 28, no. 5, pp. 1095–1108, 2017.

[21] N. Riche *et al.*, "Rare2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis," *Signal Process. Image Commun.*, vol. 28, no. 6, pp. 642–658, 2013.

[22] T. Judd *et al.*, "Learning to predict where humans look," in *2009 IEEE 12th Int. Conf. Comput. Vis.*, 2009, pp. 2106–2113.

[23] B. Schauerte and R. Stiefelhagen, "Quaternion-Based Spectral Saliency Detection for Eye Fixation Prediction," in *Proc. 12th Eur. Conf. Comput. Vis.*, A. Fitzgibbon *et al.*, Eds., 2012, pp. 116–129.

[24] A. Torralba *et al.*, "Contextual Guidance of Eye Movements and Attention in Real-World Scenes: The Role of Global Features in Object Search." *Psychol. Rev.*, vol. 113, pp. 766–786, 11 2006.

[25] X. Hou, J. Harel, and C. Koch, "Image Signature: Highlighting Sparse Salient Regions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 194–201, 2012.

[26] M. Cornia *et al.*, "A deep multi-level network for saliency prediction," in *2016 23rd Int. Conf. Pattern Recognit.*, 2016, pp. 3488–3493.

[27] W. Wang and J. Shen, "Deep Visual Attention Prediction," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2368–2378, 2018.

[28] M. Cornia *et al.*, "Predicting Human Eye Fixations via an LSTM-Based Saliency Attentive Model," *IEEE Trans. Image Process.*, vol. 27, no. 10, pp. 5142–5154, 2018.

[29] S. Jia and N. D. Bruce, "EML-NET: An Expandable Multi-Layer NETwork for saliency prediction," *Image Vis. Comput.*, vol. 95, p. 103887, 2020.

[30] R. Droste, J. Jiao, and J. A. Noble, "Unified Image and Video Saliency Modeling," in *Proc. 16th Eur. Conf. Comput. Vis.*, 2020, pp. 419–435.

[31] Z. Che *et al.*, "How is Gaze Influenced by Image Transformations? Dataset and Model," *IEEE Trans. Image Process.*, vol. 29, pp. 2287–2300, 2020.

[32] A. Kroner *et al.*, "Contextual encoder–decoder network for visual saliency prediction," *Neural Netw.*, vol. 129, pp. 261–270, 2020.

[33] F. Hu and K. McGuinness, "FastSal: a Computationally Efficient Network for Visual Saliency Prediction," in *25th Int. Conf. Pattern Recognit.*, 2021, pp. 9054–9061.

[34] J. Pan *et al.*, "SalGAN: Visual Saliency Prediction with Generative Adversarial Networks," in *arXiv*, January 2017.

[35] I. Laurent and K. Christof, "Neural Mechanisms of Selective Visual Attention," *Nat. Rev. Neurosci.*, vol. 2, p. 194–203, 2001.

[36] R. G. Alexander *et al.*, "What do radiologists look for? Advances and limitations of perceptual learning in radiologic search," *J. Vis.*, vol. 20, no. 10, pp. 17–17, 10 2020.

[37] V. Jampani *et al.*, "Assessment of Computational Visual Attention Models on Medical Images," in *Proc. 8th Indian Conf. Comput. Vis., Graph. Image Process.*, ser. ICVGIP '12. New York, NY, USA: Assoc. for Comput. Machinery, 2012, pp. 1–8.

[38] G. Wen *et al.*, "Comparative study of computational visual attention models on two-dimensional medical images," *J. Med. Imag.*, vol. 4, no. 2, pp. 1–15, 2017.

[39] D. Gu *et al.*, "VINet: A Visually Interpretable Image Diagnosis Network," *IEEE Trans. Multimedia*, vol. 22, no. 7, pp. 1720–1729, 2020.

[40] X. Lu *et al.*, "Rating Image Aesthetics Using Deep Learning," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2021–2034, 2015.

[41] X. Yang, F. Li, and H. Liu, "TTL-IQA: Transitive Transfer Learning Based No-Reference Image Quality Assessment," *IEEE Trans. Multimedia*, vol. 23, pp. 4326–4340, 2021.

[42] S. Yang *et al.*, "A Dilated Inception Network for Visual Saliency Prediction," *IEEE Trans. Multimedia*, vol. 22, no. 8, pp. 2163–2176, 2020.

[43] M. Raghu *et al.*, *Transfusion: Understanding Transfer Learning for Medical Imaging*. Red Hook, NY, USA: Curran Associates Inc., 2019, p. 3347–3357.

[44] A. van Opbroek *et al.*, "Transfer Learning Improves Supervised Image Segmentation Across Imaging Protocols," *IEEE Trans. Med. Imag.*, vol. 34, no. 5, pp. 1018–1030, 2015.

[45] J. Deng *et al.*, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[46] B. Zhou *et al.*, "Places: A 10 Million Image Database for Scene Recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1452–1464, 2018.

[47] A. Esteva *et al.*, "A guide to deep learning in healthcare," *Nat. Med.*, vol. 25, p. 24–29, 2019.

[48] L. Lévêque *et al.*, "A statistical evaluation of eye-tracking data of screening mammography: Effects of expertise and experience on image reading," *Signal Process.: Image Commun.*, vol. 78, pp. 86–93, 2019.

[49] L. Lévêque *et al.*, "State of the Art: Eye-Tracking Studies in Medical Imaging," *IEEE Access*, vol. 6, pp. 37 023–37 034, 2018.

[50] C. Privitera and L. Stark, "Algorithms for defining visual regions-of-interest: comparison with eye fixations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 9, pp. 970–982, 2000.

[51] H. Liu and I. Heynderickx, "Visual Attention in Objective Image Quality Assessment: Based on Eye-Tracking Data," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 7, pp. 971–982, 2011.

[52] T. Judd, F. Durand, and A. Torralba, "A Benchmark of Computational Models of Saliency to Predict Human Fixations," MIT Computer Science and Artificial Intelligence Lab, Cambridge, MA, USA, Tech. Rep. MIT-CSAIL-TR-2012-001, 01 2012.

[53] M. Kümmerer *et al.*, "MIT/Tübingen Saliency Benchmark," https://saliency.tuebingen.ai/.

[54] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 185–207, 2013.

[55] Z. Bylinskii *et al.*, "What Do Different Evaluation Metrics Tell Us About Saliency Models?" *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 3, pp. 740–757, 2019.

[56] X. Yang, F. Li, and H. Liu, "A Measurement for Distortion Induced Saliency Variation in Natural Images," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–14, 2021.

[57] S. Liu and W. Deng, "Very deep convolutional neural network based image classification using small training sample size," in *2015 3rd IAPR Asian Conf. Pattern Recognit.*, 2015, pp. 730–734.

[58] K. He *et al.*, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[59] L.-C. Chen *et al.*, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, 2018.

[60] M. Sandler *et al.*, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in *2018 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4510–4520.

[61] I. Goodfellow *et al.*, "Generative adversarial nets," in *Advances Neural Inf. Process. Syst.*, Z. Ghahramani *et al.*, Eds., vol. 27. Curran Associates, Inc., 2014, p. 2672–2680.

[62] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Med. Image Comput. Computer-Assisted Intervention – MICCAI 2015*, N. Navab *et al.*, Eds. Cham: Springer Int. Publishing, 2015, pp. 234–241.

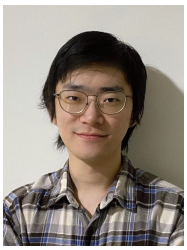
[63] M. Jiang *et al.*, "SALICON: Saliency in Context," in *2015 IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1072–1080.

[64] Z. Liu *et al.*, "A ConvNet for the 2020s," in *2022 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, June 2022, pp. 11 976–11 986.

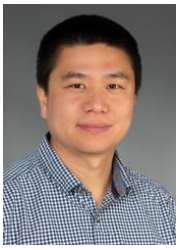
[65] J. Wang *et al.*, "Deep High-Resolution Representation Learning for Visual Recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, 2021.

[66] X. Dong *et al.*, "CSWin Transformer: A General Vision Transformer Backbone With Cross-Shaped Windows," in *2022 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, June 2022, pp. 12 124–12 134.

[67] J. Lou *et al.*, "TranSalNet: Towards perceptually relevant visual saliency prediction," *Neurocomputing*, vol. 494, pp. 455–467, 2022.
[68] M. Kümmerer, M. Bethge, and T. S. A. Wallis, "DeepGaze III: Modeling free-viewing human scanpaths with deep learning," *J. Vis.*, vol. 22, no. 5, pp. 7–7, 04 2022.
[69] M. Assens *et al.*, "Saltinet: Scan-path prediction on 360 degree images using saliency volumes," in *2017 IEEE Int. Conf. Comput. Vis. Workshops*, 2017, pp. 2331–2338.
[70] X. Chen, M. Jiang, and Q. Zhao, "Predicting human scanpaths in visual question answering," in *2021 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 10 871–10 880.
[71] M. A. Kerkouri *et al.*, "Salypath: A Deep-Based Architecture For Visual Attention Prediction," in *2021 IEEE Int. Conf. Image Process.*, 2021, pp. 1464–1468.
[72] M. Tliba *et al.*, "Self supervised scanpath prediction framework for painting images," in *2022 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2022, pp. 1538–1547.
[73] C. Sun *et al.*, "Revisiting unreasonable effectiveness of data in deep learning era," in *2017 IEEE Int. Conf. Comput. Vis.*, 2017, pp. 843–852.
[74] Y.-S. Lin, P.-H. Huang, and Y.-Y. Chen, "Deep Learning-Based Hepatocellular Carcinoma Histopathology Image Classification: Accuracy Versus Training Dataset Size," *IEEE Access*, vol. 9, pp. 33 144–33 157, 2021.
[75] S. Wang *et al.*, "Follow My Eye: Using Gaze to Supervise Computer-Aided Diagnosis," *IEEE Trans. Med. Imaging*, 2022.



Jianxun Lou received the B.Eng. from Central South University, Changsha, China, in 2018 and the M.S. degree from Cardiff University, Cardiff, UK, in 2020. He is now pursuing his Ph.D. degree at the School of Computer Science and Informatics, Cardiff University, Cardiff, UK.



Hanhe Lin received his Ph.D. at the Department of Information Science, University of Otago, New Zealand in 2016. From 2016 to 2021, he was a postdoc at the Department of Computer and Information Science at the University of Konstanz, Germany, where he was working on project A05 (visual quality assessment) of SFB-TRR 161, funded by the German Research Foundation (DFG). Currently, he is Lecturer in Computing at University of Dundee, UK. His research interests include image processing, computer vision, machine learning, deep learning, and visual quality assessment. He serves as a member of technical program committee or a reviewer in a number of conferences/journals, e.g. QoMEX, IEEE TPAMI/TIP.

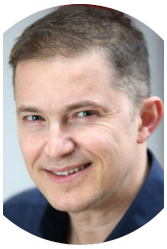
Dr Philippa Young, MA, MB BChir, MRCP, FRCR was born and educated in UK, obtaining medical qualification from University of Cambridge in 1991 and subsequently Membership of the Royal College of Physicians London in 1994. This was followed by diagnostic radiology training in Wales obtaining Fellowship of the Royal College of Radiologists in 1997. Dr Young had an initial post as a Consultant Radiologist in Taunton and Somerset NHS Trust from 1999 – 2006, and then a move back to Cardiff, Wales ensued in 2006 in order to subspecialise in breast radiology. Currently, Dr Young is a Consultant Breast Radiologist, Clinical Lead at Breast Test Wales and Breast Radiology Lead at Cardiff and Vale University Health Board. Dr Young has a keen interest in research related to breast cancer diagnosis and treatment including with use of new technologies, such as AI and ablation treatment technologies.

Dr Richard White received degrees in medical science and medicine from the University of St Andrews and the University of Manchester. He underwent radiology training in Dundee, Scotland, during which time he was conferred with Fellowship of the Royal College of Radiologists. He took up a consultant post at the University Hospital of Wales in 2013 and is currently a consultant vascular and interventional radiologist, with research interests in vascular and interventional radiology, innovations and AI. He is also the radiology R&D lead and a registered Clinical Radiation Expert with the Health Research Authority.

Zelei Yang was born in Beijing, China in 1990. He received his MBBS at University College London Medical School (London, United Kingdom) in 2015 with an intercalated BSc in Neuroscience. He is currently an interventional radiologist at University Hospital of Wales, Cardiff and is an Honorary Lecturer at Cardiff University.

Susan Shelmerdine completed her medical and specialist radiology training in London, UK. After this, she completed two paediatric radiology fellowships at Great Ormond Street Hospital (2014-15) and The Hospital for Sick Children, Toronto (2015-16). She successfully defended her PhD at UCL, London in 2020 and is currently an NIHR funded post-doctoral academic radiologist based at Great Ormond Street Hospital, London and the UCL GOS Institute of Child Health. She is a member of the Royal College of Radiologists AI committee and also chairs the European Society of Paediatric Radiology AI taskforce.

David Marshall is a Professor in the School of Computer Science and Informatics, Cardiff University, UK. He received his BSc in 1986 from University College,Cardiff in 1986 and obtained his PhD "Three Dimensional Inspection of Manufactured Objects" from there in 1990. His research interests include human facial analysis, high dimensional subspace analysis, audio/video image processing, and computational music. He has published over 150 papers and one book in these research areas. He is a member of the British Machine Vision Association. http://users.cs.cf.ac.uk/Dave.Marshall/



Prof Emiliano Spezi was born in Pesaro (Italy), he obtained both is Bachelor (Physics) and specialisation degree (Medical Physics) from the University of Bologna (Bologna, Italy). He obtained his PhD from the University of Wales College of Medicine (Cardiff, UK). After a decade of work in the National Health Service as Research Clinical Scientist, Prof Spezi moved to Cardiff University School of Engineering where is led the Medical Engineering Research Group. He is current Director of Research for the School of Engineering. Prof Spezi is a UK registered clinical scientist (Health and Care Professions Council, HCPC), a chartered member of the Institute of Physics (IoP) and a Fellow of the Institute of Physics and Engineering in Medicine (IPEM).

Marco Palombo received his Ph.D. at the Department of Physics, Sapienza University of Rome, Italy in 2014. From 2014 to 2016, he was a postdoc at the Molecular Imaging Research Centre (MIRCen) at CEA in France and from 2016 to 2020 at the Department of Computer and Information Science at the University of Konstanz, Germany, where he was working on developing computational methods for non-invasive in vivo microstructure imaging through Magnetic Resonance Imaging (MRI). Currently, he is Senior Lecturer in Microstructure Imaging at Cardiff University, UK. His research interests include biophysical and computational modeling, medical image processing and analysis, machine learning, deep learning, Bayesian inference. He serves as a member of technical program committee and reviewer in a number of conferences/journals, e.g ISMRM, MICCAI, Neuroimage, Brain, and he is a UK Research and Innovation Future Leaders Fellow.

Hantao Liu received the Ph.D. degree from the Delft University of Technology, Delft, The Netherlands in 2011. He is currently an Associate Professor with the School of Computer Science and Informatics, Cardiff University, Cardiff, U.K. He is an Associate Editor of IEEE Transactions on Circuits and Systems for Video Technology and IEEE Signal Processing Letters.